Investigating the Correlation Between Forest Carbon Stocks and Surface Temperature Changes

1 Introduction

This project aims to understand how surface temperature changes over time relate to shifts in global and regional forest carbon stocks. There are no clear answers but patterns might help us predict what will happen next or make decisions based on what is to come.

2 Research Question

Is there a correlation between changes in forest carbon stocks and surface temperature changes over the decades globally or regionally?

3 Data Sources

3.1 Datasource 1: Forest and Carbon Dataset

- **Metadata URL:** IMF Forest Data
- **Why this dataset:** Understanding the function of forests in sequestering carbon and their impact on climate change requires having access to comprehensive and precise information on forest areas and carbon stocks, which is provided by this dataset.
- **Source of data:** International Monetary Fund (IMF)
- **What data it contain:** The dataset includes **ObjectId**, which is a unique identifier for each record; **Country**, which is the name of the country; **ISO2**, which is the ISO 3166-1 alpha-2 country code; **ISO3**, which is the ISO 3166-1 alpha-3 country code; **Indicator**, which is the specific indicator being measured (forest area, carbon stocks); **Unit**, which is the unit of measurement for the indicator; **Source**, which is the source of the data; **CTS_Code**, **CTS_Name**, and **CTS_Full_Descriptor** additional metadata pertaining to the data classification; and **F1992** to **F2020** annual data for the specific indicator from 1992 to 2020.
- **Data Structure:** From 1992 to 2020 the data set has columns for countries and years among other indicators like forest area or carbon stock but here you see it in CSV format with those columns.
- **Data Quality:** The dataset is sourced from the IMF, covering very close to thirty years, making it mostly accurate, complete, and consistent.

3.2 Datasource 2: Annual Surface Temperature Change Dataset

- **Metadata URL:** IMF Surface Temperature Data
- **Why this dataset:** it offers estimates of variations in the mean surface temperature across a number of decades, the Annual Surface Temperature Change Dataset was chosen.
- **Source of data:** International Monetary Fund (IMF)
- **What data it contain:** The dataset consists of **ObjectId**, which is a unique identifier for each record; **Country**, which is the name of the country; **ISO2**, which is the ISO 3166-1 alpha-2 country code; **ISO3**, which is the ISO 3166-1 alpha-3 country code; **Indicator**, which is the specific indicator being measured (temperature change); **Unit**, which is the unit of measurement for the indicator; **Source**, which is the source of the data; **CTS_Code**, **CTS_Name**, and **CTS_Full_Descriptor** additional metadata pertaining to the classification of the data; and **F1961** to **F2022** annual data for the specific indicator from 1961 to 2022.
- **Data Structure:** The dataset, which contains annual data from 1961 to 2022, is in CSV format and has columns for the nation, year, and temperature change in degrees Celsius.
- **Data Quality:** The IMF provided the largely reliable, consistent, and comprehensive dataset that offers a long-term view of temperature variations.

3.3 Licenses

The two datasets can be used and distributed as long as correct citation and acknowledgment are given, thanks to open-data licenses. IMF Data Terms of Use.

3.4 Fulfilling License Obligations

In order to follow the guidelines set out in using these data sources, it is necessary that all project outputs properly attribute and reference them. The datasets will follow the mandates of IMF and must be accompanied by their sources when any data or visualization is produced from them. Therefore, we have to integrate identifiers of the IMF norms into our findings sharing and publishing processes to assure our conformance.

4 Data Pipeline

4.1 Data Pipeline Description

- The data pipeline is responsible for getting, cleaning, transforming, and archiving climate change-related datasets. It was made using Python while its tools are used to handle data manipulation as well as database storage tasks using sqlite3 and pandas. Data processing is done, and data fetching from outside sources is automated by it. This is in order to avail data for questioning or analysis in a SQLite database in the simplest mode.

4.2 Pipeline Steps

1. **Download Data:** Data is downloaded from URLs and saved locally.
2. **Data Cleaning:** Remove rows with missing values and convert data to numeric format.
3. **Data Transformation:** Calculate annual differences for temperature and carbon stocks.
4. **Data Storage:** Store the cleaned and transformed data in an SQLite database.

4.3 Data Pipeline Diagram

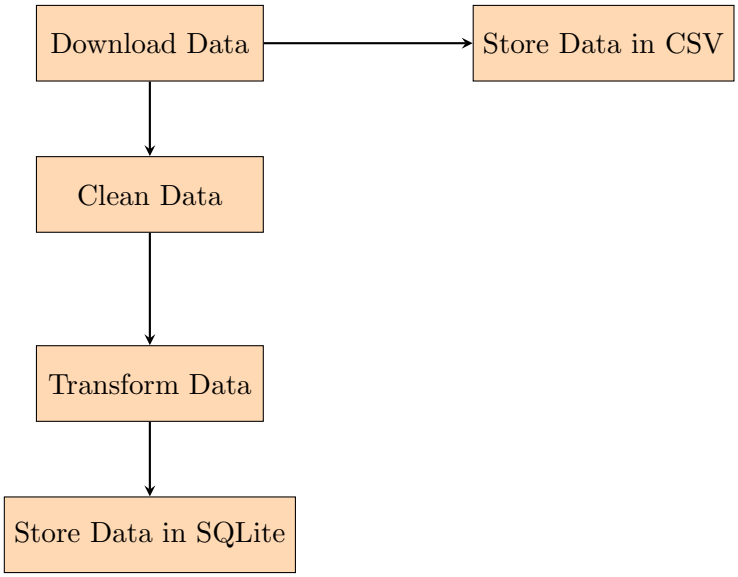

Figure 1: Data Pipeline Diagram

4.4 Transformation and Cleaning Steps

Download and Read CSVs: The pipeline starts by downloading the datasets from specified URLs and reading them into pandas DataFrames.

Filter Relevant Data:

- **Columns Removed:**
  - For the Annual Surface Temperature Change dataset: *ObjectId*, *ISO2*, *ISO3*, *Unit*, *Source*, *CTS_Code*, *CTS_Name*, and *CTS_Full_Descriptor*.
  - For the Forest and Carbon dataset: *ObjectId*, *ISO2*, *ISO3*, *Unit*, *Source*, *CTS_Code*, *CTS_Name*, and *CTS_Full_Descriptor*.
- **Reason for Removal:** The datasets were made simpler and easier to handle by eliminating these columns in order to concentrate on the main indicators of interest, namely temperature variations and carbon stocks.
- **Columns Preserved:**
  - For the Annual Surface Temperature Change dataset: *Country*, *Indicator*, and annual data columns (*F1961* to *F2022*).
  - For the Forest and Carbon dataset: *Country*, *Indicator*, and annual data columns (*F1992* to *F2020*).

Drop Missing Values: The missing value rows of both datasets have been deleted. In order for our analysis be valid, there should be consistency and completion of the dataset. Missing values were deleted so as to maintain data integrity which can mislead the findings or lead to wrong conclusions.

Convert Data Types: We used pandas to transform the columns of annual data into a numeric format. Whenever arithmetic operations or statistical analyses are to be made, the kind of data should be numeric.

Set Index: In both data systems, the index was established for the *Country* column. By organizing a data structure of the two datasets and setting the country as the index, transnational merger and comparison of data is simple.

Calculate Annual Differences: The investigators calculated the changes in carbon stores and temperature that happen from year to year. For one to understand the patterns and identify obvious shifts, it is possible for one to compare each year with its predecessor in order to keep track of yearly changes.

Store in SQLite Database: The data which was cleaned and converted was held on an SQLite database. the SQLite database is available very quickly for analysis, is easy to query fast, and helps to keep information manageable.

4.5 Problems Encountered

- **Missing values:** drop rows with NaNs.
- **Data type inconsistencies:** Confirmed that numerical columns are in correct format.

4.6 Error Handling

The pipeline has tests that verify incoherent data and reveals issues through the execution process.

5 Result and Limitations

5.1 Output Data

Data saved in a SQLite database after cleansing and transforming constitutes the Pipeline's output data. The datasets include:

- Data on changes in the average annual surface temperature for multiple countries over a number of years is contained in this temperature change information set.
- Data on forests and carbon, with carbon stocks in forests across several years for different nations.
- Analyzed yearly changes in temperature alterations and carbon reserves, thereby enabling year on year trend analysis.

5.2 Data Structure and Quality

Data Structure:

- **Format:** The data is stored in tables within an SQLite database.
- **Tables:**
  - *Annual_Surface_Temperature_Change:* Contains columns for *Country*, *Indicator*, and annual temperature change data from 1961 to 2022.
  - *Forest_and_Carbon:* Contains columns for *Country*, *Indicator*, and annual carbon stock data from 1992 to 2020.
  - *Temp_Change_Diff:* Contains the annual differences in temperature change for each country.
  - *Carbon_Stocks_Diff:* Contains the annual differences in carbon stocks for each country.

Data Quality:

- **Completeness:** The dataset is mostly intact, excluding some rows that did not include all the needed data items, so that accuracy and trustworthiness are maintained.
- **Consistency:** The dataset underwent a filtering process to remove any anomalies while ensuring that its structure is maintained after transforming numbers into meaningful digits suitable for usage in any calculation.
- **Accuracy:** Data from the reliable organization known as International Monetary Fund is what has been used thus signifying utterly high levels of accuracy and trustworthiness.

5.3 Data Format

Data Format: The cleaned and transformed data is stored in an SQLite database.
Reason for Choosing SQLite:

- **Efficiency:** SQLite is a suitable option for working with large datasets because it has efficient storage and querying capabilities.
- **Portability:** SQLite databases are contained completely, hence easily transportable or can be shared.
- **Ease of Use:** Python's `sqlite3` library makes it simple to use SQLite thereby easy integration with pandas DataFrames.

5.4 Critical Reflection and Potential Issues

Reflection on Data:

- **Strengths:**
  - The pipeline cleans and converts the data to make it consistent and suitable for analysis purposes
  - By storing the data in an SQLite database, one can query and analyze it efficiently.
- **Potential Issues:**
  - **Data Completeness:** When you delete rows that have no data it will make sure that your resultant data set will be consistent, however you are highly likely going to drop some very big points that are valuable for analysis thus making it not comprehensive.
  - **Linear Assumptions:** Annual differences are computed based on the assumption that changes between years will be linear excluding more intricate patterns or unusual data.
  - **Geographical and Temporal Coverage:** There are several countries and many years represented in the databases; nonetheless, scarce information exists for certain regions or particular periods, which might complicate the identification of trends and correlations.
  - **Analysis Limitations:** stating the inadequacies of annual difference computations linear assumptions in clear cut words.
  - **Future Adaptations:** It is important to highlight that it may be necessary to modify pipelines so as to accommodate future changes in the sources of data.